Proceedings of the 2nd International Conference on Mathematical Physics and Computational Simulation
DOI: 10.54254/2753-8818/41/2024CH0187

Review on VSLAM based on deep learning

Xin Shao

School of Control Science and Engineering, Shandong University, Shandong, China

201518220103@mail.sdu.edu.cn

Abstract. Visual simultaneous localization and mapping technology (VSLAM) provides a theoretical basis for the operation of unmanned equipment such as autonomous vehicles and sweeping robots in unfamiliar environments. Although traditional VSLAM systems have achieved great success after long-term development, it is still difficult to maintain good performance in challenging environments. Deep learning, as a newly developed technology in the field of vision in recent years, has shown outstanding advantages in image processing. Combining deep learning with VSLAM is a hot topic. Deep learning can help traditional VSLAM systems improve the lack of scale information in dynamic environments by improving the performance of traditional VSLAM in depth estimation, pose estimation, and closed loop detection. It can not only reduce the scale of the network model but also improve the accuracy of trajectory estimation. Specifically, in terms of the fusion of VSLAM method flow and deep learning, many researchers have proposed deep learning fusion methods based on visual odometry, loop detection and mapping. This work studies the trend and combination of VSLAM with deep learning algorithms, hoping to provide help for the real autonomy of future mobile robots, and finally puts forward prospects for the development of VSLAM.

Keywords: Visual simultaneous localization and mapping technology, deep Learning, end-to-end.

1. Introduction

Visual simultaneous localization and mapping (SLAM) has been an increasingly popular field of study in recent years. There are solutions based on lidar and sonar, and there are also solutions based on visual sensors mainly cameras. The former sensors are expensive and bulky, while the latter are lightweight, portable and low-cost, being widely used in the industry. VSLAM uses visual sensors to perceive the surrounding environment, build maps of complex three-dimensional spaces and achieve autonomous navigation. In domains like intelligent robotics, autonomous vehicles, drones, unmanned aerial vehicles, augmented reality (AR), and virtual reality (VR), this VSLAM technology is crucial. Unmanned vehicles in smart car factories can automatically pick and match auto parts and cooperate with the information system of the production line to achieve fully automated production. Rescue robots and underwater vehicles in complex working environments (such as electromagnetic interference and failure of GPS positioning systems) can achieve long-distance autonomous cruising, tunnel detection and deep-water rescue tasks through VSLAM technology. In addition, emerging technologies AR and VR can achieve interaction between virtual and reality. The three-dimensional map reconstructed by VSLAM can accurately render virtual images in the geometric position of the real scene, making the overall

DOI: 10.54254/2753-8818/41/2024CH0187

virtual space look more real. With the development of these fields, more novel methods and technologies will emerge in VSLAM, and VSLAM technology has become a field worthy of active research [1].

Visual odometry (VO) and loop closure principles serve as the foundation for VSLAM, which adheres to the front-end, back-end, loop detection, and map-drawing architecture of classic SLAM algorithms. By analyzing the variations between various video frames, the front-end determines the camera stance and composition of the surrounding surroundings, which is generally achieved by feature-based methods and direct methods. Due to the limited scope of the inter-frame estimate, which only takes into account two consecutive frames, there is inevitably a margin of error in the motion between each pair of images. Repetitive transmission of the error estimated between successive frames leads to error buildup and trajectory deviation. So, in order to reduce the accumulated mistakes, it is necessary to implement back-end optimization and loop detection. The front-end processing method and the matching job requirements subsequently generate a map [2].

Convolutional neural networks are extensively utilised in image recognition to extract image information, making deep learning algorithms more prevalent in this sector. The feature extraction technique is highly efficient and robust. An input layer, a hidden layer, and an output layer are the typical components of a fully functional neural network. The training methods of neural networks are generally divided into supervised, semi-supervised, and unsupervised. The supervised method uses data with labeled information to train the network, the unsupervised method provides unlabeled raw data to the network for training, and the semi-supervised method is between the two, using both labeled and unlabeled data to train the network [3].

With the rapid development of deep learning and some urgent problems in VSLAM, the fusion method of deep learning and VSLAM has become a challenge for researchers. Many literatures only describe the methods from the perspective of combining deep learning with VSLAM modules. For example, Liu Ruijun et al. introduced the combination of deep learning and VSLAM from the perspectives of odometer and closed loop detection and compared it with traditional methods, but did not outline the combination of deep learning and VSLAM from a holistic perspective [4]. This paper summarizes the latest VSLAM methods based on deep learning in recent years by outlining three methods of integrating deep learning models into traditional VSLAM systems: auxiliary modules based on deep learning, replacement modules based on deep learning, and using end-to-end neural networks to replace the overall VSLAM architecture. It can help relevant personnel better understand the current research progress and future development direction of VSLAM based on deep learning.

2. Overview of VSLAM technology

2.1. VSLAM principles

VSLAM technology comprises four essential components: visual odometry, optimization, loop closure detection, and mapping. Front-end visual odometry entails extracting distinctive characteristics from sequences of images and comparing them over frames to determine the incremental movement of the camera's position, resulting in real-time positioning. However, it is susceptible to the gradual accumulation of errors over time, leading to drift. Back-end optimization minimizes the discrepancy between the anticipated and observed feature positions within a specific time frame to reduce accumulated drift. This process is known as pose optimization. Loop closure detection identifies previously visited areas upon repeat visits. It then imposes limitations between the current and former positions to prevent any deviation. The mapping module integrates visual input and optimum poses to progressively construct a map of the unknown area [5]. Figure 1 depicts the standard sequence of tasks and the interrelationship between these components.

DOI: 10.54254/2753-8818/41/2024CH0187

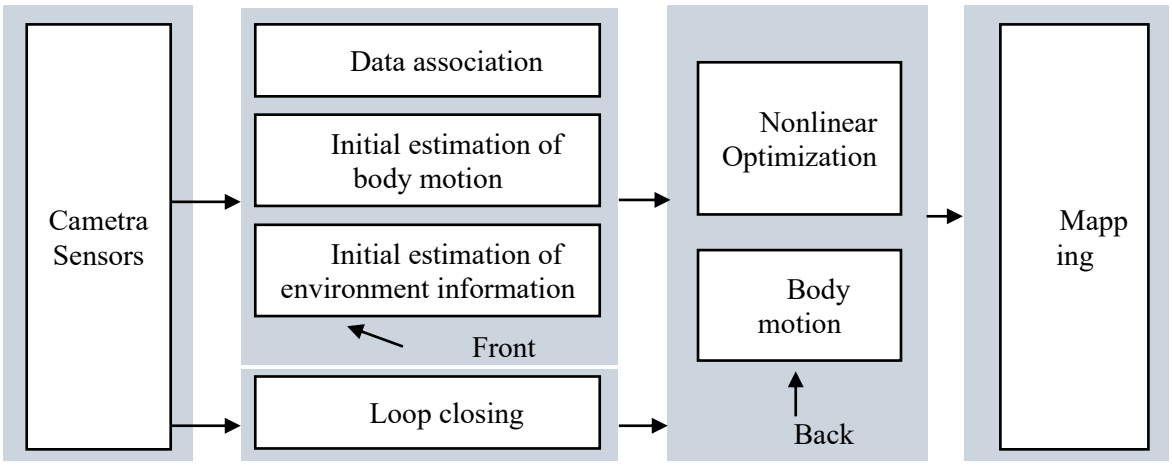

Figure 1. Basic principles of VSLAM.

2.2. Front-end visual odometry

There are two primary techniques for visual odometry: feature-based and direct. The feature point approach involves recognising the pixel differences between consecutive frames of an image in order to establish the link between picture features and calculate the relative motion between the camera and the surroundings. The feature-based approach is commonly used in visual odometry. The recent ORB-SLAM3 algorithm can be implemented by using information captured by monocular cameras, binocular cameras, RGB-D cameras, and inertial measurement unit (IMU) sensors. Compared with other algorithms, it has higher robustness, accuracy, and versatility [6]. However, the feature method performs poorly in the absence of obvious texture and when the pixel difference is small. The direct method calculates the relative motion by comparing the photometric difference between the previous and next frames of the image. It can work in areas with unclear textures, but it does not involve the global features of the image, resulting in poor closed-loop detection. In general, the challenges faced by visual odometry include lighting changes, motion blur, occlusion, and dynamic objects in the surrounding environment.

2.3. Back-end optimization

The back-end mainly uses filtering methods and nonlinear optimization methods to process and optimize the noisy data obtained from the front-end to obtain more accurate motion trajectories and spatial point positions. Filtering techniques, such as the extended Kalman filter, continuously update the estimated position at the present time by merging the motion dynamics and observing the state at the previous time step. Because the memory space occupied by the algorithm grows as the square of the state, it performs well in small spaces, but its application in large scenes is limited. The optimization method is based on the idea of graph optimization and uses all states to estimate the current situation. Although filtering methods are computationally efficient, smoothing methods improve accuracy at the expense of higher computational costs.

2.4. Loop detection

During the movement of the VSLAM system, there will be cumulative errors between the estimated pose and the environmental position. The loop detection module can identify scenes that appear repeatedly during the movement and use this recognition result to correct the map to ensure the global consistency of the map. The loop detection algorithm can effectively eliminate the cumulative error.

The primary approach for loop identification is to use the bag of words model to extract local information from the image and construct a word list consisting of k words. The scene's visuals can be represented as k-dimensional vectors based on the word list. The vector's value can then be utilised to ascertain if distinct photographs depict the same scene.

DOI: 10.54254/2753-8818/41/2024CH0187

2.5. Mapping

According to different front-end processing methods and different task requirements, it is necessary to construct maps with corresponding forms and complexity, which can not only accurately describe environmental features, but also reduce the complexity of the map while ensuring accuracy [1]. Based on varying dimensions, map representation can be categorized as two-dimensional or three-dimensional. Two-dimensional maps can be categorised into three types: geometric maps, grid maps, and topological maps. Geometric maps utilise a limited number of landmarks, such as points, line segments, and curves, to represent the characteristics of the scene environment. The grid map divides the environment into many equal-sized grids and provides a probability value to indicate the presence of an object in each grid. Each grid unit can be classified into one of three states: occupied, idle, or unknown. These states are used to differentiate between areas that can be traversed and areas that are obstructed. The topological map uses the connection lines between nodes to form a topological structure diagram to represent the scene, where the nodes are locations in the actual environment, and the connection lines between nodes represent the relationship between different locations.

Among three-dimensional maps, point cloud maps are the most widely used maps. Although point cloud maps retain detailed information about the original environment, point cloud maps are generally large in scale, and many details that are not required for many tasks take up a lot of space. An octree map, commonly referred to as a three-dimensional grid map, can be created using the octree structure. Compared with a two-dimensional grid map, an octree map is more effective in describing the environment, has less ambiguity, and saves a lot of space compared to a point cloud map. However, the corresponding computational complexity is large, so it is difficult to search and plan a real-time path. In addition, according to the specific task requirements and the front-end processing methods, different types of maps include feature maps, euclidean signed distance fields (ESDF) maps, truncated signed distance fields (TSDF) maps, semantic maps, etc.

3. VSLAM algorithm based on deep learning

Since 2010, deep learning and reinforcement learning have been actively combined with VSLAM. There are three prevalent approaches to combination: the creation of auxiliary modules using deep learning, the creation of deep learning modules, and the substitution of the entire architecture with end-to-end deep neural networks.

3.1. Deep learning algorithms

At present, the methods of monocular depth estimation using machine learning can be divided into two types, namely, the method of combining traditional machine learning with image geometric features and the method of monocular depth estimation using a convolutional neural network [7]. The former uses depth clues in the image, such as linear perspective, focus, defocus, atmospheric scattering, shadow, etc., to construct parameter equations such as Markov random field and conditional random field for training [7]. This method often does not meet the needs of actual scenes and has low prediction accuracy. Or the method based on similarity search searches for similar images that have appeared in a known data set. The limitations of the data set lead to the low generalization ability of this method, which is only applicable to specific scenes. At the same time, the retrieval time is long and cannot meet real-time requirements. The latter refers to a system that relies on deep learning. It use convolutional neural networks that have been trained with large amounts of data to create comprehensive image depth information. Two main types of deep learning methods exist: supervised learning and unsupervised learning. Supervised learning necessitates a substantial level of monitoring as a training component. The training accuracy is high, but the difficulty lies in the acquisition of real depth in the data set. Unsupervised learning does not require real depth values to train the network. It uses binocular image pairs or video sequences as input and realizes supervision during network training by designing a reasonable loss function.

DOI: 10.54254/2753-8818/41/2024CH0187

3.2. Module analysis based on deep learning

By substituting one of the four modules of standard VSLAM, namely front-end, back-end, loop detection, and map drawing, with an independently trained neural network, the overall performance of VSLAM can be enhanced. This is referred to as an auxiliary method that relies on deep learning.

LIFT-SLAM relies on the process of optimizing feature extraction [8]. The system utilizes the learning invariant feature transform (LIFT) to extract features from pictures. The conventional VSLAM pipeline, based on ORB-SLAM, then incorporates these features for applications involving monocular cameras. Using learned features at the front end of the VSLAM system can provide advantages by enabling the acquisition of denser and more accurate matches. Furthermore, the uniform distribution of these characteristics throughout the image results in a more consistent motion estimation. Several studies have confirmed the resilience and high efficiency of this VSLAM algorithm. Utilizing VO sequence photos for training deep neural networks (DNN) can result in the extraction of more effective task-specific features. Transfer learning can enhance the performance of the overall system on cross-datasets by fine-tuning these networks using VO/VSLAM datasets. Furthermore, a method has been successfully developed to dynamically modify the matching threshold based on the number of outliers throughout the execution of the visual odometry (VO) pipeline. This method enables the removal of the predetermined value of the matching threshold without the need for dataset adjustment.

TransPoseNet is an optimization technique that relies on pose recognition [9]. The suggested method efficiently detects geometric information in low-light photos, unaffected by the indistinct texture caused by inadequate illumination. The fundamental structure involves conducting initial identification, followed by subsequent identification, which is accomplished via deep learning and keypoint-based geometric alignment. The initial stage of identification entails simultaneously performing depth completion and posture regression to mitigate the visual alterations caused by occlusion in the depth image. During the refinement stage, the ICP alignment framework uses keypoints instead of full depth image points to improve localization efficiency. Weakly supervised pose regression identifies keypoints on the depth feature map. The authors proved that their method works better than common keypoint detectors like SIFT and SURF by using the 7-Scenes dataset, which is made up of a collection of RGB-D frames.

DRM-SLAM is an optimization technique that relies on map reconstruction [10]. The use of a Convolutional Neural Network (CNN) that is specifically developed using the ResNet architecture enables the accomplishment of real-time dense and accurate depth estimation as well as scene reconstruction. The deep fusion method, which is based on the deep reconstruction model, makes the most of the sparse depth samples that ORB-SLAM generates and the depth map that CNN infers to reconstruct the image in a dense and accurate way.

PlaceNet is an optimization technique that relies on the closure of loops detection [11]. PlaceNet is an innovative numerous scale deep autoencoder network that incorporates a semantic fusion layer to improve scene comprehension. The primary concept behind PlaceNet is to acquire knowledge about areas in a dynamic environment that should be disregarded due to the presence of moving items. In other words, it aims to prevent distractions caused by dynamic objects and instead concentrate on significant features within the scene. PlaceNet is trained to identify dynamic objects in a scene by acquiring knowledge of a grayscale semantic map that indicates the positions of both stationary and mobile objects inside an image. PlaceNet produces deep features that are aware of the meaning of the environment and are resistant to changes in scale and dynamics.

DeepSLAM is a recently developed visual SLAM framework that relies on end-to-end learning [12]. The system takes a series of individual color stereo photos as input and simultaneously learns the robot's position and the three-dimensional representation of the surrounding environment in a complete, unsupervised manner. This system's exclusive use of RGB input during testing enables its application in a variety of environments, including both indoor and outdoor ones.

DOI: 10.54254/2753-8818/41/2024CH0187

4. Conclusion

VSLAM, a fast-advancing scientific discipline, has garnered significant interest from numerous academics who are involved in the development and utilization of deep learning models. Recent advancements in deep learning have significantly enhanced many phases involved in VSLAM processing, including as data processing, posture estimation, trajectory estimation, mapping, and loop closure. This paper primarily organizes fundamental information on visual SLAM and deep learning, and presents the current application state of visual SLAM with deep learning in four key areas: visual odometer, backend optimization, loop detection, and mapping module. Finally, a typical case of end-to-end neural networks in VSLAM is mentioned. It can be found that end-to-end learning can directly optimize all VSLAM modules at the same time, providing a model that is more resilient to noise and uncertainty. End-to-end deep neural networks show significant potential in improving the performance of VSLAM algorithms. The basic structure for these architectures is on self-supervised learning and reinforcement learning, which enable adaptability in actual dynamic environments. By combining traditional methods like Kalman filters or Savitzky-Golay filters with end-to-end deep models, enhanced outcomes can be achieved. End-to-end DNN are very flexible and can be used in many different fields, including surgery, figuring out the pose of a drone, controlling automated underwater vehicles, navigating drones, and mapping altitude. Constructing a comprehensive learning framework is a complex task, since it needs meticulous management of the connections between modules in a discernible manner to enable learning through backpropagation. Deep learning models possess inherent constraints. As an illustration, they are unable to analyze inertial data along with color, depth, and LiDAR data. Consequently, future endeavors will require thorough and comprehensive investigation.

In generally, deep learning models present possibilities for processing visual data in real-time and with high efficiency, although there are challenges in integrating data from various sensor types.

References

[1] Zhang Yao, Wu Yiquan & Chen Huixian. (2023). Research progress of visual simultaneous localization and mapping based on deep learning. Journal of instruments and meters (07), 214-241. The doi: 10.19650 / j.carol carroll nki cjsi. J2311081.

[2] Favorskaya, M. N. (2023). Deep learning for visual SLAM: The state-of-the-art and future trends. Electronics, 12(9), 2006. doi:https://doi.org/10.3390/electronics12092006

[3] Sun H. (2023). Master's Degree in VSLAM system based on monocular depth Estimation (Dissertation, Hangzhou Dianzi University). Master of http://link.cnki.net.https.gzlib.proxy.chaoxing.com/doi/10.27075/d.cnki.ghzdc.2023.000809 Doi: 10.27075 /, dc nki. GHZDC. 2023.000809.

[4] Liu Ruijun, Wang Shangxiang, Zhang Chen, et al. Visual SLAM based on deep learning review [J]. Journal of system simulation, 2020, 32 (7): 1244-1256. The DOI: 10.16182 / j.issn1004731x. Joss. 19 - vr0466.

[5] Chen, S.; Zhou, B.; Jiang, C.; Xue, W.; Li, Q. A lidar/visual slam backend with loop closure detection and graph optimization. Remote Sens. 2021, 13, 2720.

[6] Campos, C., Elvira, R., Gómez Rodríguez, J.,J., Montiel, J. M. M., & Tardós, J.,D. (2021). ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. Ithaca: doi:https://doi.org/10.1109/TRO.2021.3075644

[7] Shang Guangtao, Chen Weifeng, Ji Aihong, et al. VSLAM review based on neural network [J]. Journal of nanjing information engineering university, 2024 (03) : 352-363. The DOI: 10.13878 / j.carol carroll nki jnuist. 20220420001.

[8] Li, Q.; Cao, R.; Zhu, J.; Fu, H.; Zhou, B.; Fang, X.; Jia, S.; Zhang, S.; Liu, K.; Li, Q. Learn then match: A fast coarse-to-fine depth image-based indoor localization framework for dark environments via deep learning and keypoint-based geometry alignment. ISPRS J. Photogramm. Remote Sens. 2023, 195, 169–177. [CrossRef]

[9] Bruno, H. M. S., & Colombini, E. L. (2021). LIFT-SLAM: A deep-learning feature-based monocular visual SLAM method. Ithaca: doi:https://doi.org/10.1016/j.neucom.2021.05.027

DOI: 10.54254/2753-8818/41/2024CH0187

[10] Ye, X.; Ji, X.; Sun, B.; Chen, S.; Wang, Z.; Li, H. DRM-SLAM: Towards dense reconstruction of monocular SLAM with scene depth fusion. Neurocomputing 2020, 396, 76–91

[11] Hussein Osman, Nevin Darwish, AbdElMoniem Bayoumi, PlaceNet: A multi-scale semantic-aware model for visual loop closure detection, Engineering Applications of Artificial Intelligence, Volume 119, 2023, 105797, ISSN 0952-1976, https://doi.org/10.1016/j.engappai.2022.105797.

[12] R. Li, S. Wang and D. Gu, "DeepSLAM: A Robust Monocular SLAM System With Unsupervised Deep Learning," in IEEE Transactions on Industrial Electronics, vol. 68, no. 4, pp. 3577-3587, April 2021, doi: 10.1109/TIE.2020.2982096.